DOCUMENT RESUME

ED 169 120 TM 008 581

TITLE	Guidelines for Reviewing the Technical Quality of Examinations for Granting College Credit.
INSTITUTION	American Council on Education, Washington, D.C. Office on Educational Credit.
PUB DATE	1 Oct 78
NOTE	10p.
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Credits; *Equivalency Tests; *Evaluation Criteria; Higher Education; *Standards; *Test Reviews; *Test Selection

ABSTRACT

A team of technical experts should review the technical quality of proficiency examinations considered for assessing learning and granting college credit. The team should recommend approval or disapproval of the test; and if it is approved, recommend the level of test performance required in order to award credit. Information supplied by the test publisher may be helpful in making these decisions. The American Psychological Association's Standards for Educational and Psychological Tests represent a primary basis for judging tests. A number of other criteria should be considered. The skill or ability domains measured by the test should be clearly stated, and the rationale for item selection should be indicated. Item analysis data should indicate how well the domain has been sampled. Test items should be well-written and previously tested. Reliability data should pertain to test use in awarding credit by examination; cutting scores and norms should be appropriate for the institution using the test. The test should be free of known bias. The publisher should provide instructions for testing, scoring, time limits, and test security. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

American Council on Education
Office on Educational Credit
Credit-By-Examination Test Review Program

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Douglas R. Whitney

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

GUIDELINES FOR REVIEWING THE TECHNICAL QUALITY OF EXAMINATIONS FOR GRANTING COLLEGE CREDIT

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED169120

The Task

The technical review team is to prepare a report that (1) critically reviews the technical quality of the test, (2) recommends that the test be approved or disapproved for the purposes of assessing learning and awarding college credit, and (3) recommends, if approval seems warranted, the level of performance on the examination that should be required in order for credit to be awarded. The report form used is the Report of Technical Review Team. This completed report should represent a comprehensive test review and reflect the consensus of the team. Individual team members are free to file minority reports or comments on issues about which consensus could not be reached. A copy of all reports will be sent to the test publisher.

The team should have a set of materials supplied by the test publisher and a completed Publisher's Checklist. The checklist will serve as an index to help locate data and other information needed to answer questions about the technical quality of the test being reviewed.

In making these judgements the following points should be kept in mind:

1. Subject matter specialists may be requested to make independent evaluations of the content of the test.

2. "Textbook perfection" is seldom attainable and is not required. An appreciation for the practical problems which test developers encounter in editing items, pre-testing and norming such tests and a reasonable tolerance for less-than-perfect solutions are expected.

3. Successful performance on approved tests will result in college credit being awarded to many students. Some of the decisions to award or deny credit will be in error because the tests are imperfect. Tests should be approved only when it is believed that the percentage of erroneous decisions due to the inadequacies of the test is likely to be considerably smaller than the percentage of good decisions.

TM008 581

General Considerations

Throughout the review process, professional judgment is to be exercised. The tests should be judged from the point of view of whether they merit approval for use in granting college credit.

Tests used for granting college credit include a variety of types, and some may not follow the traditional model of psychological trait measurement (e.g., tests built following latent trait theory, performance tests, simulation tests, objectives-based tests, etc.). As a consequence, traditional statistical indices that have been used to judge test quality may not always be appropriate. Nevertheless, the claim that the test is an instrument that will provide valid and reliable information for decisions about granting college credit must be supported by both empirical data and logical consideration.

The Standards for Educational and Psychological Tests (American Psychological Association, 1974) represent a primary basis for judging the tests. The Standards, however, are broad based and rooted in traditional psychometric principles, not all of which may be applicable in the usual way to every type of test that might be used to grant college credit for skills and knowledge already attained. Though some specifics of the Standards may not apply in the same way to all tests to be reviewed, there still remains the basic principle that the test publisher must provide evidence of the test's validity and reliability and support any claims made for its use.

The end result of the review is a recommendation for either approval or disapproval of the test for the purpose of granting college credit. Conditional or partial approval cannot be made by the review team. If disapproval is recommended, the reasons for this decision are to be documented in the Report of Technical Review Team. Also, the conditions that must be met in order for the disapproval recommendation to be changed to an approval recommendation should be listed.

Special Considerations

Although the principles outlined in the Standards for Educational and Psychological Tests form the primary basis for the review, there are certain considerations that should be specifically taken into account when forming a judgement about a particular test.

1. There should be a clear statement defining the domain(s) of content, skills, and/or developed abilities that the test is intended to sample.

This domain statement should be sufficiently clear to allow the test user (who is assumed to be knowledgeable about the subject matter but not necessarily about psychometrics) to decide (a) whether the test publisher's domain definition matches the user's own course, program or area of study, (b) whether the content emphasis of the test matches the intended domain, and (c) whether the test items represent an adequate sample from the domain. The statement should be sufficiently clear to permit a panel of content speicalists to understand what the examination is intended to measure.

2. The domain specification should indicate the rationale for choosing the kinds of tasks that compose the test.

Tests may contain a variety of stimulus materials designed to elicit examinee responses. Selection of particular stimulus materials should have been made on some reasoned basis rather than haphazardly or for the sake of convenience. For example, if the examination consists of both essay questions and recognition items, the unique contribution that each is intended to make should be clearly specified.

3. There should be enough information about the individual test tasks (items) to allow qualified reviewers to judge whether the total collection constitutes an adequate sample of the pertinent domain(s).

This information may result from logical and/or statistical analyse Examples of logical analysis include classifying items according to the test specifications; describing how items were written, reviewed, and revised; and outlining model responses for essay or performance items. Examples of statistical analysis include summaries of ratings of items by a representative panel of content specialists, item analysis data from samples representing appropriate reference populations, and data summarizing scores on individual essay questions or ratings obtained by using the scoring keys developed for the questions or responses of examinees sampled from appropriate reference populations.

4. Test items should be expressed in clear and precise language (contain no major flaws) and should have been pretested before assembling the operational form of the test.

Professional judgment will be required to determine whether these aspects of test construction should be a cause for concern about recommending test approval. For example, security restrictions on the use of the test may preclude large scale pretesting of essay questions; if so, a more modest tryout may be sufficient to locate faults in questions or to improve the scoring procedures. As a general rule, objective items should be pretested in order to locate flawed items and to suggest ways to improve item quality. The focus of an item analysis should not simply be on computing statistical indices. There should be a clear rationale as to why certain items were selected for inclusion while others were discarded after the tryout phase.

5. Information about the reliability of the test scores for relevant examinee population(s) should be related specifically to issues in the use of the test for awarding credit by examination.

Determining the effects on examinee scores of both different content samples (as would be involved in multiple forms of the test) seems relevant. Tests having alternate forms ought to provide evidence of equivalence relating to both the absolute level and the relative level of the score, since absolute score interpretations will usually be made in this setting. Score stability over meaningful periods of time may also be relevant. In addition, scorer reliability data is important for essay questions and other assessment methods requiring judges to rate examinee responses, performances, or behaviors.

Concerns about reliability should relate to the "credit/no-credit" decision that will result from the use of the test. Some publishers will recommend a specific cut-score* (or several cut-scores if multiple classifications are to be made). Errors in classification are of concern in these instances. Enough information should be provided for the test user and the review team to judge the degree of misclassification likely to occur when the test developer's recommended cut-scores are used.

Errors of misclassification are a concern even if the developer does not recommend specific cut-score(s) but, instead, encourages the user to establish cut-scores at the local institution. The recommended procedures for accomplishing this should be judged according to the likelihood that they will lead users to establish cut-scores having a high degree of classification accuracy.

*The term "minimum credit awarding score" is frequently used in these guidelines instead of "cut-score." The minimum credit awarding score is the lowest score an examinee can attain and still be awarded credit based on the test. For more details refer to the discussion of "Minimum Credit Awarding Score" on pages 4-5 of the Guidelines for Reviewing the Content of Examinations for Granting College Credit.

6. The procedures for setting minimum credit awarding scores should be based on considered judgement about the use of the test and the nature of institutional and examinee populations for which its use is suggested.

There should be evidence that any minimum credit awarding score recommended by the publisher is reasonable and appropriate. What is considered "reasonable and appropriate" will depend on the particular test and the educational institution(s) at which it is to be used. The recommendation of a minimum credit awarding score by subject matter specialists, without empirical support on the outcomes expected from the use of this particular numerical value, is usually insufficient evidence. Such a professional judgment may result in a minimum score that is too high; consequently, few persons who would be considered successful by the usual classroom criteria would be granted credit based on their test performance. Similarly, procedures using rigid statistical criteria may not be appropriate. The statistically optimal level of performance may fall outside of the ability range of those persons attending most institutions at which the test is intended to be used.

These guidelines do not seek to dictate the particular method to be used to set the minimum credit awarding scores. Rather, the team is asked to judge whether the score recommended by the test publisher is reasonable and appropriate for use in granting college credit.

If the test publisher places the responsibility for establishing the minimum credit awarding score on the user, recommended procedures for the user to follow in setting this score should be given. The team should judge the reasonableness of these recommended procedures.

The American Council on Education urges each institution to examine the the test materials, collect the necessary data, and make its own final judgment on the minimum credit awarding score to be used. However, some institutions, in

addition to conducting their own test reviews, may seek the guidance of external validating sources. The Council reviews will provide those institutions with recommendations for examinations which have been approved.

If a test receives technical approval, the technical review team will evaluate the test publisher's recommendation(s) on the minimum credit awarding score. The team either concurs or does not concur with the recommendation. If the team does not concur, then it makes an alternative recommendation. Both the alternative recommendation and the test publisher's recommendation will be published. When the publisher makes no recommendation at all, the Council may ask the content review team for a recommendation, based on its best subjective judgement.

7. The test should be free of irrelevant sources of difficulty and have no known bias against specific groups.

Test items or tasks should be free from irrelevant sources of difficulty which may increase the magnitude of errors of measurement, contribute to ambiguity, or be offensive or unfair to specific groups of persons. Although there is no widely accepted agreement concerning the particular evidence needed or analytical method to be used in evaluating for test bias, reviewers should be sensitive to this concern*. The influence of factors which are known to contribute to the differential performance of certain groups, and which are irrelevant to the knowledge and skills to be certified by the test, should be at a minimum. For example, it may be true that many of those seeking college credit by examination are less test-wise than those who have attained the same knowledge or skill through course work. The influence of such factors should be evaluated.

8. There should be clear and detailed instructions for administering and scoring the test and time limits should be ample to permit most examinees to finish the test.

---

*Empirical evidence should be provided if possible.

The importance of specified and standardized test directions and administration procedures in establishing the comparability of scores from different administrations is well known. Time limits should be generous enough so that examinee work rate does not become an irrelevant source of difficulty. Tests may be speeded (or partially speeded) only when speed of performance is a demonstrably relevant part of the competency being assessed. Where speed is not relevant, reviewers should seek evidence that the test scores are not influenced unduly by the time limits.

A scoring key should be provided for use in the review. For essay questions and performance ratings, the procedures for scoring responses should be explicitly defined and should be judged for their technical quality and resulting scorer reliability. Procedures for training scorers or raters should be examined.

9. Norms for the test should be based on a representative sample from the appropriate population(s) and should be accurate within a statistically acceptable margin of error.

For some tests and for some credit awarding decisions, normative-referenced information is appropriate and desirable. The test publisher should clearly describe the characteristics of the appropriate population and explain why this population is appropriate. The publisher should demonstrate that the sample was drawn from the appropriate population(s) in a way that insures representativeness. Although the size of the sample selected for the development of the norms is important to obtain statistical precision, the number of cases composing the norming sample is less important than are the degree of representativeness and the appropriateness of the population sampled. Since the test publisher should provide norms that are appropriate for the uses that are claimed for the test, the intended uses should be clearly articulated and the test's norms should be judged in light of them.

10. The test publisher should make provisions for adequate test security.

Because the tasks on most tests will be considered samples from the domain defining the knowledge or skill area for which college credit will be given, adequate provisions must be made for test security. This is necessary to insure that the tasks sampled do reflect the domain intended. Test security concerns extend not only to the administration and use of the test, but also to its initial developmental stages. Generally, examinees should not have access to the test questions before taking the test. If examinees are permitted to re-take the examination, alternate forms should be available for this purpose.